

UNCLASSIFIED

Defense Technical Information Center
Compilation Part Notice

ADP014054

TITLE: Archiving and Preservation in Electronic Libraries

DISTRIBUTION: Approved for public release, distribution unlimited

Availability: Hard copy only.

This paper is part of the following report:

TITLE: Electronic Information Management for PfP Nations [La gestion
electronique des informations pour les pays du PfP]

To order the complete compilation report, use: ADA415655

The component part is provided here to allow users access to individually authored sections of proceedings, annals, symposia, etc. However, the component should be considered within the context of the overall compilation report and not as a stand-alone technical report.

The following component part numbers comprise the compilation report:

ADP014048 thru ADP014057

UNCLASSIFIED

Archiving and Preservation in Electronic Libraries

Gail Hodge

Information International Associates, Inc.
312 Walnut Place
Havertown, Pennsylvania 19083, USA

Abstract

The rapid growth in the creation and dissemination of electronic information has emphasized the digital environment's speed and ease of dissemination with little regard for its long-term preservation and access. To some extent, electronic libraries, that is those libraries that are moving toward provision of materials in electronic form, have been swept up in this attitude as well. But, electronic information is fragile in ways that traditional paper-based information is not. Electronic information is more easily corrupted or altered, intentionally or unintentionally, without the ability to recognize that the corruption has occurred. Digital storage media have unknown life spans. Some formats, such as multimedia, are so closely linked to the software and hardware technologies that they cannot be used outside these proprietary environments. Aggravating this situation is the fact that the time between creation and preservation is shrinking, because technological advances are occurring so quickly.

While there is a tradition of stewardship, best practices, and stakeholder roles that has long been institutionalized in the print environment, many of these traditions are inadequate, inappropriate or not well known among the stakeholders in the digital environment. Creators of electronic resources are able to bypass the traditional publication, dissemination and announcement processes that have been part of the path from creation to archiving and preservation in the print environment. Publishers and librarians who traditionally managed this process must now look to computer scientists to develop systems that support these activities. Digital libraries may be the responsibility of computer scientists who do not necessarily bring skills in content management, organization and preservation. Best practices and policies are needed that satisfy both the requirements of the digital environment and the economic interests of the various stakeholder groups.

Electronic information is information that is born digital or that has its primary version in digital form. Electronic information includes a variety of object types, such as electronic journals, e-books, databases, data sets, reference works, and web sites. These are the types of information that electronic libraries are trying to manage and preserve.

The Open Archival Information System (OAIS) Reference Model provides a framework for discussing the key areas that impact on digital preservation -- the creation of the electronic information, the acquisition of and policies surrounding the archiving of resources, preservation formats, preservation planning including issues of migration versus emulation, and long-term access to the archive's contents.

Many projects, worldwide, have contributed to the growing collection of best practices and standards. The numerous stakeholder groups involved in preservation of electronic resources, including creators (authors), publishers, librarians and archivists, and third-party service providers, are working more closely to build a cohesive and sustainable response to the issues. An issue of continuing stakeholder interest is the economic model(s) that will provide ongoing support to electronic preservation.

Despite the remaining issues, local institutions managing electronic libraries can become involved. They are encouraged to monitor developments and projects in the field, to raise awareness of the need for preservation within their institutions, to consider preservation and long-term access issues when negotiating licenses for electronic resources, and to look for opportunities to begin small projects at the local level.

1.0 Background

1.1 Definition of Terms

Several terms will be used throughout this lecture. They are defined here. In some cases, these definitions are for consistency within the presentation and are not indicative of general consensus within the community.

Born digital – materials that are created in bits and bytes rather than being digitized from paper or other analog medium

Digital archiving – storing the digital information for long term preservation

Digital preservation – keeping the bits and bytes safe and unaltered for a long period of time

Digitization – converting materials in non-digital form (analog) such as paper, to digital form

Emulation – running old products by recreating the environment of the old hardware and software without actually using the old hardware and software

Long-term access – the ability to use a preserved object long after its initial preservation

Migration – moving a digital product from one version of a program, operating system or hardware environment to another over time

Recapturing – copying the content from the original resource again in order to ensure that changes made to the resources are incorporated in the archival version

Refreshing – moving a digital object to a new instance of the same media, retaining the same operating system and hardware environment

1.2 Outline of Major Projects

I've selected several major projects in digital archiving as examples. (For a more complete list, I recommend the PADI (Preserving Access to Digital Information) Web site from the National Library of Australia (NLA 2002).) I will briefly describe these since they are used throughout the remainder of the lecture.

CAMiLEON, (Creative Archiving at Michigan and Leeds: Emulating the Old on the New) a joint project of the University of Michigan and the University of Leeds, is conducting analysis and testing to determine if emulation is a viable technical strategy for preservation. (University of Michigan)

Cedars (CURL Exemplars in Digital Archiving) is sponsored by the Joint Information Systems Committee in the UK. It was established to determine the feasibility of distributed digital archives. The first implementation included the three institutions in the Consortium of University Research Libraries. In the last two years, Cedars has included several other test sites. (Cedars)

ERPANET (Electronic Resources Preservation and Access Network) is a new project funded by the European Commission to provide a knowledge base and advice to all sectors on issues of archiving and preservation of electronic resources. (ERPANET)

EVA is a project of the National Library of Finland at the University of Helsinki. It uses a series of automatic tools including robots, harvesters, and metadata creation tools to support its goal of capturing electronic network publications of Finland. (Lounamaa and Salonharju 1999)

InterPARES (International Research on Permanent Authentic Records in Electronic Systems) is a global project among seven archiving institutions, including regional consortia for Asia and Europe. The project's goals are to develop best practices related to the creation, preservation and long-term access to *authentic* electronic records. (InterPARES)

Kulturaw3 is a project of the Royal Library of Sweden. Its goal is to capture the cultural heritage that is being published via the Internet. Unfortunately, this project has been stopped due to the lack of deposit legislation for digital materials in Sweden. (National Library of Sweden)

LOCKSS (Lots of Copies Keep Stuff Safe) is a project of the Stanford University Library, its publishing arm, HighWire Press, and several other libraries to develop a system for redundant archives. Its major contribution is an infrastructure for keeping redundant archives synchronized. (LOCKSS)

JSTOR, originally funded by the Andrew J. Mellon Foundation, is now a non-profit organization that archives back issues of journals for publishers by digitizing them. It is just beginning to deal with current journal issues that are in electronic form. (JSTOR)

NEDLIB (the Network European Deposit Libraries) was funded by the European Union. It included eight libraries and numerous publishers and other organizations. This project was completed early in 2001. The major output was a model for incorporating archives into integrated library systems, work on metadata, early adoption and testing of the Open Archival Information System Reference Model, and testing of emulation strategies. The output of this project is available in a series of reports. The major findings are being incorporated into operational systems at the British Library and the Dutch National Library. (NEDLIB)

OCLC Digital Archive is a service of OCLC that grew out of its electronic journals' project. In this service OCLC acts as a trusted third party archive receiving deposits of electronic journals into its repository. It provides several levels of access (continuous or just in case) and controls access rights so that a library can access only the issues equating to the period for which it had a license. (OCLC Digital Archive)

PANDORA (Preserving and Accessing Networked DOcumentary Resources of Australia), a project of the National Library of Australia, captures the Web-based cultural heritage of Australia. It involves capturing content, creating metadata, and making arrangements with rights holders. A federated approach is envisioned that includes the libraries in all the Australian states. (PANDORA)

OCLC/RLG Preservation Metadata Working Group is a joint project that also includes members from other major projects include Cedars and the Digital Preservation Coalition. The major effort at this time is on establishing a standard element set for preservation metadata. (OCLC 2000, 2001)

2.0 A framework for archiving and preservation

It is valuable to discuss archiving and preservation within a framework. The framework I've chosen is provided by a reference model, which is being used extensively throughout the digital preservation community. The Open Archival Information System Reference Model (CCSDS 2001) provides high level data and functional models and a consistent terminology for discussing preservation. The reference model was originally developed by the Consultative Committee on Space Data Systems to support the archiving of data among the major space agencies. However, it has become the de facto standard for the development of digital archives. It is used by most major projects including those in Australia, the United Kingdom, the Netherlands, and the United States. The OAIS Reference Model is a draft standard of the International Standards Organization and is expected to be formally balloted in 2002.

In its simplest form the OAIS looks like this (Fig. 1):

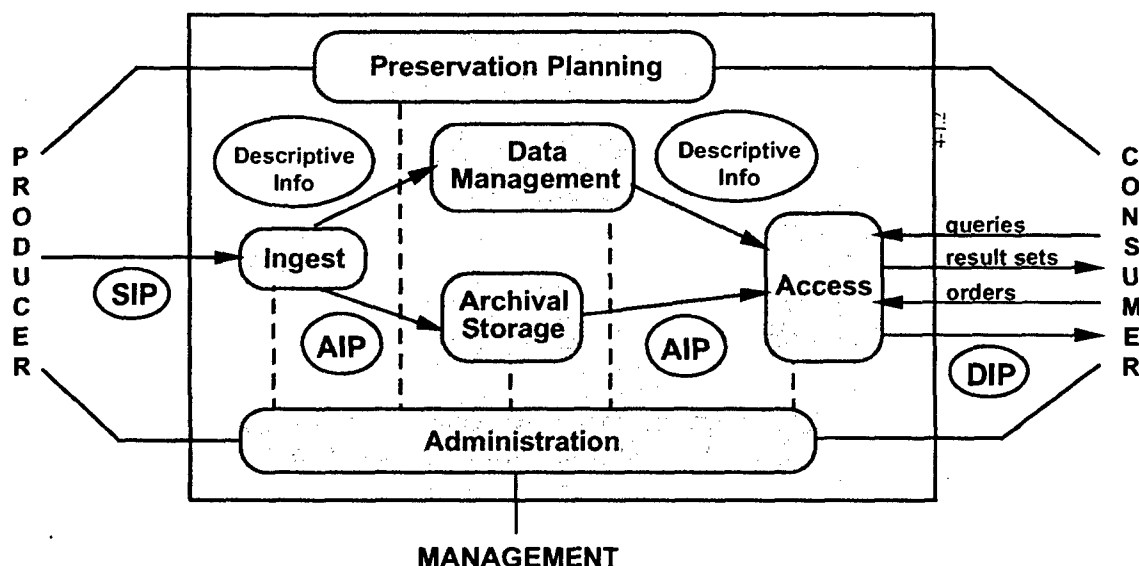


Fig. 1. Open Archival Information System

Source: Consultative Committee on Space Data Systems (used with permission)

SIP – Submission Information Packet (what is submitted or acquired from the producer)

AIP – Archival Information Packet (the object that is archived)

DIP – Dissemination Information Packet (the object that is distributed based on access requests)

Descriptive Info – metadata

2.1 Production and creation of electronic information

Archiving begins outside the purview of the archive with the producer or the creator of the electronic resource. This is where long-term archiving and preservation must begin. Information that is born digital may be lost if the producer is unaware of the importance of archiving. Practices used when electronic information is produced will impact the ease with which the information can be digitally archived and preserved.

Several key practices are emerging involving the producers of electronic information. First, the preservation and archiving process is made more efficient when attention is paid to issues of consistency, format, standardization and metadata description before the material is considered for archiving. By limiting the format and layout of certain types of resources, archiving is made easier. This is, of course, easier for a small institution or a single company to enforce than for a national archive or library. In the latter cases, they are faced with a wide variety of formats that must be ingested, managed and preserved.

In the case of more formally published materials, such as electronic journals, efforts are underway to determine standards that will facilitate archiving. The Andrew J. Mellon Foundation has funded a study of the electronic journal mark-up practices of several publishers. The study concluded that a single SGML document type definition (DTD) or XML schema can be developed to support the archiving of electronic journals from different subject disciplines and from different publishers with some loss of special features (Inera, Inc. 2002). Such standardization is considered key to efficient archiving of electronic journals by third-party vendors.

In the case of less formally published material such as web sites, the creator may be involved in assessing the long-term value of the information. In lieu of other assessment factors, the creator's estimate of the long-term value of the information may be a good indication of the value that will be placed on it by members of its designated community or audience in the future. The Preservation Office at the National Library of Medicine has implemented a "permanence rating system" (Byrnes 2001). The rating is based on three factors: integrity,

persistent location, and constancy of content. These factors have been combined into a scheme that can be applied to any electronic resource. At the present time, the ratings are being applied to NLM's internal Web sites, and guidelines have been developed to assist creators in assigning the ratings to their sites. This information will be used to manage the ongoing preservation activities and to alert users about a Web site's long-term stability.

Another aspect of the creator's involvement in preservation is the creation of metadata. The best practice is for metadata to be created prior to incorporation into the archive, i.e., at the producer stage. However, most of the metadata continues to be created "by hand" and after-the-fact. Unfortunately, metadata creation is not sufficiently incorporated into the tools for the creation of most objects to rely on the creation process alone. However, as standards groups and vendors move to incorporate XML and other architectures into software products, such as word processors, the creation of metadata should become easier and more automatic.

2.2 Ingest: Acquisition and collection development

Now moving into the functions to be performed by the archive itself. The first is acquisition and collection development. This is the stage in which the created object is "incorporated" physically or virtually into the archive. In the terminology of the reference model, this is called "Ingest".

There are two main aspects to the acquisition of electronic information for archiving – collection policies and gathering procedures.

2.2.1 Collection policies

Just as in the paper environment, there is more material that could be archived than there are resources with which to accomplish it. Guidelines are needed to tailor the collection policies to the needs of a particular organization and to establish the boundaries in a situation where the responsibility for archiving among the stakeholders is still unregulated. The collection policies answer questions such as what should be archived, what is the extent of a digital object, should the links that point from the object to be archived to other objects also be archived, and how often should the content of an archived site be recaptured?

2.2.1.1 Selecting what to archive

In the network environment where any individual can be a publisher, the publishing process does not always provide the screening and selection at the manuscript stage on which traditional archiving policy has relied. Therefore, libraries are left with a larger burden of selection responsibility to ensure that publications of lasting cultural and research value are preserved (NLC 1998).

The scope of NLA's PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) Project is to preserve Australian Internet publishing. The NLA has formulated guidelines for the *Selection of Online Australian Publications Intended for Preservation by the National Library of Australia* (NLA): Scholarly publications of national significance and those of current and long term research value are archived comprehensively. Other items are archived on a selective basis "to provide a broad cultural snapshot of how Australians are using the Internet to disseminate information, express opinions, lobby, and publish their creative work." The National Library of Canada has written similar guidelines (NLC 1998). The broadest guidelines for Collection Management are provided in a draft document from the Cedars Project (Weinberger 2000). The most comprehensive analysis of such guidelines is in the Digital Preservation Handbook, which is based on the combined lessons learned of all the major projects (Beagrie and Jones 2001).

Even the Internet Archive (Internet Archive), which considers the capture of the entire contents of the Internet as its mandate, has established limitations. The sites selected do not include those that are "off-limits," because they are behind firewalls, require passwords to access, or are hidden within Web-accessible databases, and those that require payment.

The major lesson from the selection guidelines is the importance of creating such a document in order to set the scope, develop a common understanding, and inform the users now and in the future what they can expect from the archive.

2.2.1.2 Determining extent

Once the site has been selected for inclusion, it is necessary to address the issue of extent. What is the extent or the boundary of a particular digital work, especially when selecting complex Web sites? Is it a "home page" and all the pages underneath it, or are the units to be archived (and cataloged) at a more specific level?

The PANDORA (NLA/PANDORA) project in Australia evaluates both the higher and lower site pages to determine which pages form a cohesive unit for purposes of preservation, cataloging, and long-term access. While preference is given to breaking down large sites into components, the final decisions about extent depend upon which pages cluster together to form a stand alone unit that conveys valuable information. Each individual component must meet PANDORA's initial selection guidelines.

2.2.1.3 Archiving links

The extensive use of links in electronic publications raises the question of whether these links and their contents should be archived along with the original site. The answer to this question by any particular project will depend on the purpose of the archiving, the anticipated stability of the links, and the degree to which they contribute to the overall information value of the site.

Most organizations archive the URLs (Uniform Resource Locators) or other identifiers for the links and not the content of the linked pages, citing problems with the instability of links. Some projects have established variants on this approach. For example, PANDORA's decision to archive the content of linked objects is based on its selection guidelines; the content of the linked site is captured only if it meets the same selection criteria as other sites. The National Library of Canada captures the text of a linked object as long as it is on the same server as the object that is being archived, because these intra-server links have proven to be more stable than external links. The American Institute of Physics (AIP) points to the content of a linked reference if it is an item in AIP's archive of publications or supplemental material.

Elsevier, which is currently involved in an archiving project with the Yale University Library funded by the Andrew J. Mellon Foundation, cites a technology-related problem as the main reason it does not archive links (Hunter 2002). Elsevier's links are created on the fly, so there is no URL or live page to capture. Similar problems exist when trying to capture pages that are active server pages or those that are created out of a database, portal system, or content management system.

The American Astronomical Society (AAS) has perhaps the most comprehensive approach to the archiving of links. The AAS maintains all links to documents and supporting materials based on collaboration among the various astronomical societies, researchers, universities and government agencies involved in this specific domain. Each organization archives its own publications, retaining all links and access to the full text of all other links. Within this specific domain, the contents of all linked objects are archived. In the future, similar levels of cooperation may be achieved in other subject domains or by publisher collaborations such as CrossRef.

2.2.1.4 Recapturing the archived contents

In cases where the site selected for archiving is updated periodically, recapturing the object is necessary. This would be the case for an electronic journal that publishes each article online as it becomes available or for a preprint service that allows the author to modify the content of the preprint as it proceeds through the review process.

When making decisions about recapturing the content of an archived site, a balance must be struck between the completeness and currency of the archive and the burden on the system resources. PANDORA allocates a gathering schedule to each "publication" in its automatic harvesting program. The options include on/off, weekly, monthly, quarterly, half-yearly, every nine months, or annually. The selection is dependent on the degree of change expected and the overall stability of the site. When making decisions about recapturing the content, the EVA Project (Lounamaa and Salonharju 1999) at the University of Helsinki considers the burden on its system resources and the burden on the sites (because of the activities of its robots) from which the content would be recaptured.

2.2.2 Gathering procedures

There are two general ways in which the archive acquires material. The producer can submit the material to be archived, or the archive can gather the material proactively.

In the first method, the best practices identified in the earlier section on creation become extremely important. Even within an organization, where the producer and the archive are almost one and the same organization, attention to standardization and limitations on the number of formats will have a significant impact on the ease with which submissions can be processed.

In the second approach, the archive may or may not have a formal relationship with the creator or the producer. In this gathering approach, the information to be archived may be hand-selected or harvested automatically. In the case of the NLA, sites are identified, reviewed, hand-selected, and monitored for their persistence before being captured for the archive.

In contrast, the Royal Library, the National Library of Sweden, until recently automatically acquired material by running a robot to capture sites for its Kulturaw3 project (National Library of Sweden). The harvester automatically captured sites from the .se country domain and from foreign sites with material about Sweden, such as travel information or translations of Swedish literature. While the acquisition was automatic, priority was given to periodicals, static documents, and HTML pages. Conferences, usenet groups, ftp archives, and databases were considered lower priority. Unfortunately, this project has been discontinued because of the lack of national deposit legislation for electronic materials.

2.3 Data management: Metadata for preservation

Metadata is needed to preserve the object and for users in the future to find and access it. Metadata supports organization, preservation and long-term access. In this section, I will deal primarily with metadata for preservation. Other issues surrounding metadata for description and discovery were covered in my previous lecture on Cataloging and Indexing of Electronic Resources.

Archiving and preservation require special metadata elements to track the lineage of a digital object (where it came from and how it has changed over time), to detail its physical characteristics, and to document its behavior in order to reproduce it on future technologies. Each of the major preservation projects – Cedars, PANDORA, NEDLIB, the Harvard Library Project, etc., had its own set of metadata that it considered important for preservation. In 2000, the Research Libraries Group and OCLC reviewed the various sets of preservation metadata and concluded that there is sufficient similarity among the elements that a core set of metadata for preservation could be identified (RLG 2000).

In October 2001, the Preservation Metadata Working Group developed a draft set of over 20 elements and numerous sub-elements for metadata preservation in the framework of the OAIS Reference Model (RLG 2001). They describe the content and the environment (software and operating systems) needed for the object. The plan is to achieve international consensus on this set. OCLC is already beginning to use the set as the basis for its Digital Archive and for the work that has been done with the U.S. Government Printing Office.

The discussion provided by the Preservation Metadata Working Group acknowledges that there are details about the use of the proposed elements and perhaps additional elements needed to preserve objects of various types. A recent meeting sponsored by UNESCO, the ICSU Press, the international Committee on Data and the International Council for Scientific and Technical Information raised the issue of whether the proposed preservation metadata element set is broad enough to encompass data sets. Based on this meeting, the data community plans to review the current draft and comment about any revisions necessary to broaden or clarify its scope.

2.4 Archival storage: Formats for preservation

A major issue for the archiving community is which format(s) should be used for archival storage. Should the electronic resource be transformed into a format more conducive to archiving? Is the complexity of an interactive journal necessary or should it be simplified? Should consideration be given to the re-use of

information and its enhancement or representation in more advanced access technologies in the future? Should the goal be complete replication of the electronic resource or should preservation provide a copy that is "just good enough"? (For example, Cedars has identified the concept of "significant properties," which are properties that are absolutely required in order for a user in the future to get the information value from the resource (Russell 2000).)

Of course the answer to these questions in part differ by resource type, and there is little standardization at this point. Most electronic journals, reference book, or reports use image files (TIFF), PDF, or HTML. TIFF is the most prevalent for those organizations that are involved with conversion of paper issues of journals. For example, JSTOR (JSTOR), a non-profit organization that supports both storage of current journal issues in electronic format and conversion of back issues, processes everything from paper into TIFF and then scans the TIFF image. The OCR, because it cannot achieve 100% accuracy, is used only for searching; the TIFF image is the actual delivery format that the user sees. However, this does not allow embedded references to be active hyperlinks.

SGML (Standard Generalized Mark-up Language) is used by many large publishers after years of converting publication systems from proprietary formats to SGML. The American Astrological Society (AAS) has a richly encoded SGML format that is used as the archival format from which numerous other formats, including IITML and PDF, are made (Boyce 1997).

For purely electronic documents, Adobe's PDF (Portable Document Format) is the most prevalent format. This provides a replica of the Postscript format of the document, but relies upon proprietary encoding technologies. PDF is used both for formal publications and grey literature. While PDF is increasingly accepted, concerns remain for long-term preservation and it may not be accepted as a legal depository format, because it is a proprietary format.

Preserving the "look and feel" is difficult in the text environment, but it is even more difficult in the multimedia environment, where there is a tightly coupled interplay between software, hardware and content. The University of California at San Diego has developed a model for object-based archiving that allows various levels and types of metadata with separate storage of the multimedia components in systems that are best suited to the component's data type. The UCSD work is funded by the U.S. National Archives and Records Administration and the U.S. Patent and Trademark Office.

2.5 Preservation planning: Migration and emulation

Preservation planning is the bridge between the decisions made about archival storage of the bits and bytes and issues of future access and user needs. There is no common agreement on the definition of long-term preservation, but some have defined it as being long enough to be concerned about changes in technology and changes in the user community. This may be as short as 2-10 years.

Two strategies for preservation are migration and emulation. Migration means copying the object to be archived and moving it to newer hardware and software as the technology changes. Migration is, of course, a more viable option if the organization is dealing with well-established commercial software such as Oracle or Microsoft Word. However, even in these cases migration is not guaranteed to work for all data types, and it becomes particularly unreliable if the information product has used sophisticated software features. Unfortunately, this level of standardization and ease of migration is not as readily available among technologies used in fields of study where specialized systems and instruments are used.

Emulation, a strategy that replicates the behavior of old software and hardware on new hardware and software, is being considered as an alternative to migration. There are several types of emulation. Encapsulation would store information about the behavior of the hardware/software with the object. For example, a MS Word 2000 document would be labeled as such and then metadata information would be stored with the object to indicate how to reconstruct the document at the engineering --- bits and bytes -- level. An alternative to encapsulating the behavior with every instance is to create an emulation registry that uniquely identifies the hardware and software environments and provides information on how to recreate the environment. Each instance would

point to the registry. (Rothenberg 1999; 2000). Taking emulation a step further is the idea of creating a virtual machine – a new machine that based on the information in the registry could replicate the behavior of the hardware/software of the past (Lorie 2001).

While the best practice for the foreseeable future continues to be migration, emulation has been tested with some success by the CAMiLEON Project (University of Michigan). This is a joint project between the University of Michigan and the University of Leeds to determine if emulation is a viable long-term strategy for preservation. Granger has concluded that a variety of preservation strategies and technologies should be available. Some simple objects may benefit from migration, while others that are more complex may require emulation (Holdsworth and Wheatley 2001; Granger 2000).

2.6 Access

The life cycle functions discussed so far are performed for the purpose of ensuring continuous access to the material in the archive. Successful practices must consider changes to access mechanisms, as well as rights management and security requirements over the long term.

2.6.1 Access mechanisms

While many preservation projects are concerned about the ability to provide long-term access to the electronic information as it exists today, others are interested in how they might actually improve access to current information in the future. A major reason for storing the information related to the U.S. National Library of Medicine's Profiles of Science materials in TIFF and other standardized forms, such as tagged ASCII, is so that the information can be re-purposed or enhanced. Even in its development stage, the project was able to improve the quality of the video clips by converting them to High Definition Video. The belief is that there will always be newer and better technologies, and a goal of the archive is to be able to take advantage of these advances in the future.

2.6.2 Rights management and security requirements

One of the most difficult access issues for digital archiving involves rights management. What rights does the archive have? What rights do various user groups have? What rights has the owner retained? How will the access mechanism interact with the archive's metadata to ensure that these rights are managed properly? How will access rights be updated as the material's copyright status or security level changes.

3.0 Emerging stakeholder roles

A number of stakeholders can be identified including creators/authors, publishers, libraries, archives, Internet service providers, secondary publishers, aggregators, and, of course, users (Haynes, et al. 1997; Hodge 1999; Hodge 2000). The roles these various stakeholders will play in the archiving process described above remains unclear, but there are several types of electronic information for which some patterns of responsibility are emerging.

In the early stages of the digital age, most electronic journal publishers considered the creation of an electronic archive to be the same as the internal production system. However, many publishers have since come to realize that archiving and production are not one and the same function. In some cases, they are quite antithetical.

The current environment shows a growing understanding of the need for archiving and long-term preservation among the major electronic journal publishers. This may not be the situation with smaller learned society publishers, but that may be more an issue of economics than of desire. The major electronic journal publishers such as Elsevier, Nature and Blackwell have projects underway. These projects are significant, because they bring together the publishers and the major research and national libraries that have been at the forefront of the demand for publisher attention to archiving issues, particularly in their license agreements.

Librarians and archivists, particularly those at national libraries, were early advocates of digital preservation issues. Many national libraries spearheaded initiatives and research projects without additional funds and without legislative mandates to cover digital deposit. In most cases, these projects have been instrumental in advancing the research and implementation of operational systems.

In addition to new roles for publishers and librarians/archivists, trusted third party archives are emerging. These third parties, such as the OCLC Digital Archive (OCLC Digital Archive), JSTOR (JSTOR), and PubMedCentral (NCBI), see archiving/preservation as an additional business/service opportunity.

A significant outgrowth of the OAIS Reference Model has been RLG's development of attributes of an OAIS-compliant archive (RLG 2001). A certification process is being discussed that would assure a library, publisher or other organization that a particular third-party archive meets minimal requirements for importing and exporting and basic functionality related to the other aspects of an archive.

Another significant development in the emergence of clearer stakeholder responsibilities, particularly for commercially published materials, is a January 2002 announcement on digital preservation by the International Federation of Library Associations and Institutions (IFLA) and the International Publishers Association (IPA). The draft presented for discussion highlights the importance of "born digital" materials and suggests that the appropriate place for preservation of last resort is with the national libraries. It is hoped that additional legislative/policy efforts and funding for cooperative initiatives will result from this statement and from the inclusion of digital preservation on the agendas of these two major international stakeholder organizations.

4.0 Trends and issues

The trend in archiving and preservation has moved from theoretical discussions to pragmatic projects. There are more initiatives focused on the realistic details of metadata, selection criteria, technologies and systems for archiving. While the need to raise awareness has not completely disappeared, more time is being spent on partnership development, testing and implementation.

The focus of research and development has shifted to "filling in the gaps." The National Science Foundation (NSF) and the Library of Congress have created a committee of interested federal agencies, including the National Archives and the national libraries, to identify key areas of research that could be supported by NSF and other federal grants. The major research areas identified to-date include the migration of extremely large data sets and long-term access to complex multimedia objects. The Dutch National Library and the British Library efforts have developed system specifications based on the NEDLIB Project, which ended in 2001. NEDLIB developed a data and process model for the deposit of digital materials in a national library setting by working with several European publishers and eight national libraries (van de Werf 2000; Feenstra 2000). The Deposit System for Electronic Publications (DSEP), which is based on the OAIS Reference Model, will be implemented in a system to be developed by IBM.

In addition to the trend toward pragmatic initiatives, cooperation has increased among projects and across stakeholder groups. OCLC, the UK's Digital Preservation Coalition (JISC) and RLG have been instrumental in identifying, supporting and advancing key areas of cooperation. As a real sign of maturity, the work is being "divided up". While some projects are developing operational systems, others are working in the background to achieve consensus on standards among/between projects. Unlike many standards activities in the past that have developed from local and regional practices, the work related to digital archiving is starting with the goal of international consensus.

Despite these positive trends, key issues remain. The cost of archiving and the lack of established business models that will sustain long-term preservation may prove to be significant stumbling blocks in the advancement of the cause of preservation. However, even these issues are being addressed in a pragmatic fashion. OCLC, Stanford University Libraries/HighWire Press, JSTOR, and major publishers such as Elsevier are actively dealing with questions of cost and how and who will pay for the archiving. The creation of groups such as the OCLC Digital Preservation & Co-op (OCLC Digital Preservation & Co-op) will provide venues where barriers can be identified and business models can be tested. Projects such as the archive of

Elsevier material at Yale Library (also funded by the Mellon Foundation) (Hunter 2002) will further identify archiving practices that can accommodate the needs of libraries, users and the economic requirements of producers.

Another key issue for electronic libraries is intellectual property rights. Some progress has been made in the area of legal deposit of electronic resources but much remains to be done. Several initiatives show that there is increased awareness on the part of governments to address this issue. In the UK the British Library is making great strides in its voluntary digital deposit program. The Library of Congress has received appropriations from the U.S. Congress following a study by the U.S. National Research Council (NRC 2001). The funding is to produce a plan for development of an infrastructure to support federated digital preservation for the U.S. For electronic records, the InterPARES project incorporates seven major groups, including two regional archive groups from Asia and Europe (InterPARES).

5.0 Local institutional responses

We've talked about a number of projects internationally. Many of these are on a national, regional or even global scale. However, what can a local institution do to ensure the preservation of electronic resources?

First, it is important to be aware of what is going on in this field. What are the outcomes of the major projects? How are standards being developed?

There are several sources for this information. The major projects have extensive web sites, and many like Cedars and NEDLIB have produced numerous publications, which are available from the web sites. Secondly, the PADI (NLA/PADI) site at the NLA is the major portal to digital archiving information. The Electronic Resources Preservation and Access Network (ERPANET) portal promises to provide practical information and links to experts. Newsletters such as *RLG DigiNews* from the Research Libraries Group are an excellent source of up-to-date information.

The local librarian should take every opportunity to raise awareness about the importance of digital preservation at his or her institution. When possible, be proactive in seeking funds to start small projects for preserving digital materials.

A concrete way to raise awareness is to ensure that archiving and preservation are considered when negotiating licenses for electronic resources, such as electronic journals and databases. With many national regimes for deposit of digital materials lagging behind the practical uses of these materials, it is important to address these archiving issues in license agreements. Equally, it is important to try to establish a balance between the rights of the rights holders and those of the library and users.

The major lesson is to think globally but to act locally – scaling the findings of the major global activities to the local needs.

6.0 Conclusions

A review of the cutting-edge projects shows the beginning of a body of best practices for digital archiving. The early adopters in the area of digital archiving are providing lessons that can be adopted by others in the stakeholder communities. Through the collaborative efforts of the various stakeholder groups – creators, librarians, archivists, funding sources, and publishers – a new tradition of stewardship will be developed to ensure the preservation and continued access to our intellectual heritage.

7.0 References

- Beagrie, N. and Jones, M. (2001). *Preservation management of digital materials: a handbook*. London: The British Library.
- Boyce, P. (1997, November). Costs, archiving, and the publishing process in electronic STM journals. *Against the Grain*, 9(5): 86. [Online]. Available: <http://www.aas.org/~pboyce/epubs/atg98a-2.html> [3 May 2002].

- Byrnes, M. (2000). Assigning permanence levels to NLM's electronic publications. *Presented at Information Infrastructures for Digital Preservation: A One Day Workshop, Dec. 6, 2000, York, England*. [Online]. Available: <http://www.rlg.org/events/pres-2000/infopapers.html/byrnes.html> [3 May 2002].
- CAMiLEON: Creating creative archiving at Michigan & Leeds: Emulating the old on the new. (2001). [Online]. Available: <http://www.si.umich.edu/CAMiLEON/> [3 May 2002].
- Cedars: CURL Exemplars in Digital Archives. [Online]. Available: <http://www.leeds.ac.uk/cedars/> [3 May 2002].
- Committee on an Information Technology Strategy for the Library of Congress, Computer Sciences and Telecommunications Board, National Research Council. (2001). *LC21: A Digital Strategy for the Library of Congress*. National Academy Press: Washington DC. [Online]. Available: <http://books.nap.edu/books/0309071445/html/index.html> [3 May 2002].
- Consultative Committee for Space Data Systems. (2001). Reference model for an Open Archival Information System (OAIS). Red Book CCSDS 650.0-R-2, June 2001. [Online]. Available: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html [3 May 2002].
- Digital Preservation Coalition. (2002). [Online]. Available: <http://www.jisc.ac.uk/dner/preservation/prescoalition.html> [3 May 2002].
- Electronic Resource Preservation and Access NETwork: ERPANET. [Online]. Available: <http://www.erpanet.org> [3 May 2002].
- Feenstra, B. (2000). Standards for implementation of a DSEP. NEDLIB Report Series; #4, Koninklijke Bibliotheek: Den Haag. [Online]. Available: <http://www.kb.nl/coop/nedlib/results/NEDLIBstandards.pdf> [3 May 2002].
- Granger, S. (2000). Emulation as a digital preservation strategy. *D-Lib Magazine*, 6(10). [Online]. Available: <http://www.dlib.org/dlib/october00/granger/10granger.html> [3 May 2002].
- Haynes, D., Streatfield, D., Jowett, T. and Blake, M. (1997). Responsibility for digital archiving and long term access to digital data. JISC/NPO Studies on Preservation of Electronic Materials. [Online]. Available: http://www.ukoln.ac.uk/services/papers/bl/jisc-npo67/digital_preservation.html [3 May 2002].
- Hodge, G. (1999). Digital electronic archiving: The state of the art, the state of the practice. [Online]. Available: <http://www.icsti.org/conferences.html> [3 May 2002].
- Hodge, G. (2000). Digital archiving: Bringing stakeholders and issues together: A report on the ICSTI/ICSU Press Workshop on Digital Archiving. *ICSTI Forum* 33. [Online]. Available: <http://www.icsti.org/forum/33/#Hodge> [3 May 2002].
- Holdsworth, D. and Wheatley, P. (2001). Emulation, preservation and abstraction. *RLG DigiNews*, 5 (4), Feature #2. [Online]. Available: <http://www.rlg.org/preserv/diginews/diginews5-4.html#feature2> [3 May 2002].
- Hunter, K. (2002). Yale-Elsevier Mellon Project. [Online]. Available: http://www.niso.org/presentations/hunter-ppt_01_22_02/index.htm [3 May 2002].
- Inera Inc., (2001). E-journal archive DTD feasibility study. Prepared for the Harvard University Library, Office of Information Systems E-Journal Archiving Project. Pg. 62-63. [Online]. Available: <http://www.diglib.org/preserve/hadtdfs.pdf> [3 May 2002].
- Internet Archive: Building an 'Internet Library'. (2001). [Online]. Available: <http://www.archive.org> [3 May 2002].
- InterPARES: International Research on Permanent Authentic Records in Electronic Systems. (2002). [Online]. Available: <http://www.interpares.org> www.archive.org [3 May 2002].
- JSTOR: The Scholarly Journal Archive. (2002). [Online]. Available: <http://www.jstor.org> [3 May 2002].
- Kuny, T. (1998, May). The digital dark ages? Challenges in the preservation of electronic information. *International Preservation News*, 17. [Online]. Available: <http://www.ifla.org/VI/4/news/17-98.htm>
- LOCKSS: Permanent publishing on the Web. [Online]. Available: <http://lockss.stanford.edu/index.html> [3 May 2002].
- Loric, R. (2001, June). A project on preservation of digital data. *RLG DigiNews*, 5 (3), Feature # 2. [Online]. Available: <http://www.rlg.org/preserv/diginews/diginews5-3.html#1> [3 May 2002].
- Lounamaa, K. and Salonharju, I. (1999, January). EVA-the acquisition and archiving of electronic network publications in Finland." *Tietolinja News*, 1. [Online]. Available: <http://www.lib.helsinki.fi/tietolinja/0199/evaart.html> [3 May 2002].
- National Library of Australia. Selection of online Australian publications intended for preservation by the National Library of Australia. (n.d.) [Online]. Available: <http://pandora.nla.gov.au/selectionguidelines.html> [3 May 2002].

- National Library of Canada, Electronic Collections Coordinating Group. (1998). Networked Electronic Publications Policy and Guidelines. [Online]. Available: <http://www.nlc-bnc.ca/9/8/index-e.html> [3 May 2002].
- Networked European Deposit Library: NEDLIB. (2001). [Online]. Available: <http://www.konbib.nl/nedlib> [3 May 2002].
- PADI: Preserving Access to Digital Information. (1999). [Online]. Available: <http://www.nla.gov.au/padi/> [3 May 2002].
- PANDORA. [Online]. Available: <http://pandora.nla.gov.au/index.html> [3 May 2002].
- PubMed Central: an Archive of life science journals. (2002). [Online]. Available: <http://www.pubmedcentral.nih.gov/>
- OCLC Digital Archive. (2002). [Online]. Available: <http://www.oclc.org/digitalpreservation/about/archive/> [3 May 2002].
- OCLC Digital Preservation Resources, Digital & Preservation Co-op. (2002). [Online]. Available: <http://www.oclc.org/digitalpreservation/about/co-op/> [3 May 2002].
- OCLC/RLG Working Group on Preservation Metadata. (2001, October). A recommendation for content information. [Online]. Available: <http://www.oclc.org/research/pmwg/contentinformation.pdf> [3 May 2002].
- Planning Committee of the OCLC/RLG Working Group on Preservation Metadata. (2001, January). Preservation metadata for digital objects: A review of the state of the art. [Online]. Available: http://www.oclc.org/research/pmwg/presmeta_wp.pdf [3 May 2002].
- Research Libraries Group. (2001, August). Attributes of a trusted digital repository for digital materials: Meeting the needs for research resources. [Online]. Available: <http://www.rlg.org/longterm/attributes01.pdf> [3 May 2002].
- Ross, S. (2000). Changing trains at Wigan: Digital preservation and the future of scholarship. National Preservation Office: The British Library.
- Rothenberg, J. (1999, January). Avoiding technological quicksand: Finding a viable technical foundation for digital preservation. Report to CLIR. [Online]. Available: <http://www.clir.org/pubs/reports/rothenberg/contents.html> [3 May 2002].
- Rothenberg, J. (2000, April). An experiment in using emulation to preserve digital publications. NEDLIB Report Series; 1. [Online]. Available: <http://www.kb.nl/coop/nedlib/results/NEDLIBemulation.pdf> [3 May 2002].
- Royal Library. National Library of Sweden. (n.d.) Kulturav3 – Heritage Project: Long term preservation of published electronic documents. [Online]. Available: <http://www.kb.se/ENG/kbstart.htm> [3 May 2002].
- Russell, K. (2000). Digital preservation and the Cedars Project experience. Presented at Preservation 2000: An International Conference on the Preservation and Long-Term Accessibility of Digital Materials, York, England, December 7-8, 2000. [Online]. Available: <http://www.rlg.org/events/pres-2000/russell.html> [3 May 2002].
- Seville, C. and Weinberger, E. (2000, June). Intellectual property rights lessons from the CEDARS Project for digital preservation. (Draft) [Online]. Available: <http://www.leeds.ac.uk/cedars/colman/CIW03.pdf> [3 May 2002].
- van de Werf, T. (2000). A process model: The deposit system for electronic publications," NEDLIB Report Series; # 6, Koninklijke Bibliotheek: Den Haag, [Online]. Available: <http://www.kb.nl/coop/nedlib/results/DSEPprocessmodel.pdf> [3 May 2002].
- Weinberger, E. (2000, June). Toward collection management guidance." (Draft) [Online]. Available: <http://www.leeds.ac.uk/cedars/colman/CIW02r.html> [3 May 2002].